Data Wrangling

The data for this project was gathered from different sources, which are the

- Twitter API
- Udacity website.

The data consist of three different datasets: ' df_twitter table, `df_dog_image table,` and `df_dog_ table.` Each dataset contains additional data which are related to each other.

1. Twitter API

   Access to the Twitter API privileges is given to a Twitter developer account, where access to users' tweets and favorite count, and retweet count can be collected and analyzed by a third party. The data is given in JSON format.

2. Udacity website

   - Twitter_archive.csv:

     The Udacity website provides access to the tweets from the @welovedogs Twitter account from their archives. The WeRateDogs Twitter archive contains primary data for all 5000+ of their tweets. This information is contained in a CSV file called Twitter_archive.csv.

   - Image_prediction.tsv :

     Images posted on the @welovedogs Twitter account were trained by running through a neural network to classify the breeds of dogs. The result is a table full of the top three image predictions alongside each tweet ID, image URL, and the image number corresponding to the most confident predictions.

Df_dog Table

• **Tweet_id**	unique The id number of a tweet
• **In_reply_to_status_id**	The unique id number of a tweet reply
• **in_reply_to_user_id**	The unique id of the user repiled
• **timestamp**	The timestamp of the tweet
• **Source**	The source of the tweet
• **text**	The tweet contents
• **retweeted_status_id**	Unique tweet for status retweet
• **retweeted_status_user_id**	Unique user id for status retweet
• **expanded_urls**	links
• **rating_numerator**	ratings
• **name**	Dogs name
• **doggo**	
• **pupper**	
• **floofer**	
• **puppo**	

Df_Twitter Table

Tweet_id	Unique_tweet id
date_created	Date Tweeted
retweet_count	Number of retweets
Favourite_count	Number of favorites

Df_dog_image Table

tweet_id	Unique tweet id
jpg_url	Link to image
Image number	The image number predicted as a tweet can have more than one image(up to 4)
p1	P1 is the algorithm's #1 prediction for the image in the tweet
p1_conf	p1_conf is how confident the algorithm is in its #1 prediction
p1_dog	p1_dog is whether or not the #1 prediction is a breed of dog
p2	P2 is the algorithm's #2 prediction for the image in the tweet
p2_conf	p2_conf is how confident the algorithm is in its #2 prediction
p2_dog	p2_dog is whether or not the #2 prediction is a breed of dog
p3	P3 is the algorithm's #3 prediction for the image in the tweet
p3_conf	p3_conf is how confident the algorithm is in its #3 prediction
p3_dog	p3_dog is whether or not the #3 prediction is a breed of dog